Luhn Revisited: Significant Words Language Models

Mostafa Dehghani[1] Hosein Azarbonyad[2] Jaap Kamps[1]
Djoerd Hiemstra[3] Maarten Marx[2]
[1]Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
[2]Informatics Institute, University of Amsterdam, The Netherlands
[3]University of Twente, The Netherlands
{dehghani,h.azarbonyad,kamps,maartenmarx}@uva.nl,d.hiemstra@utwente.nl

ABSTRACT

Users tend to articulate their complex information needs in only a few keywords, making underspecified statements of request the main bottleneck for retrieval effectiveness. Taking advantage of feedback information is one of the best ways to enrich the query representation, but can also lead to loss of query focus and harm performance—in particular when the initial query retrieves only little relevant information—when overfitting to accidental features of the particular observed feedback documents. Inspired by the early work of Luhn [24], we propose *significant words language models* of feedback documents that capture all, and only, the significant shared terms from feedback documents. We adjust the weights of common terms that are already well explained by the document collection as well as the weight of rare terms that are only explained by specific feedback documents, which eventually results in having only the significant terms left in the feedback model.

Our main contributions are the following. First, we present significant words language models as the effective models capturing the essential terms and their probabilities. Second, we apply the resulting models to the relevance feedback task, and see a better performance over the state-of-the-art methods. Third, we see that the estimation method is remarkably robust making the models insensitive to noisy non-relevant terms in feedback documents. Our general observation is that the significant words language models more accurately capture relevance by excluding general terms and feedback document specific terms.

Keywords: Significant Words Language Models; Relevance Feedback; Pseudo Relevance Feedback.

1. INTRODUCTION

One of the key factors affecting search quality is the fact that our queries are ultra-short statements of our complex information needs. Query expansion has been proven to be an effective technique to bring agreement between user information need and relevant documents [14]. Taking feedback information into account is a common approach for enriching the representation of queries and consequently improving retrieval performance. In True Relevance Feedback (TRF), given a set of judged documents either explicitly assessed by the user or implicitly inferred from user behavior, the system tries to enrich the query to improve the performance of the retrieval. However, feedback information is not available in most practical settings. An alternate approach is Pseudo Relevance Feedback (PRF), also called blind relevance feedback, which uses the top-ranked documents in the initial retrieved results for the feedback.

CIKM'16 , October 24 - 28, 2016, Indianapolis, IN, USA

ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983814

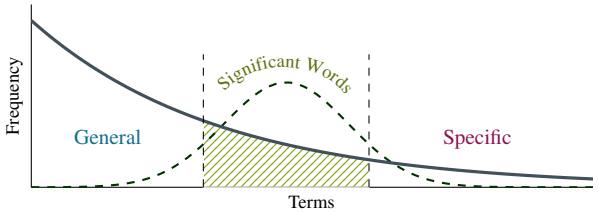


Figure 1: Establishing a set of "Significant Words" based on Luhn [24]

The main goal of feedback systems is to make use of feedback documents to estimate more accurate query models representing the notion of relevance. However, although documents in the feedback set contain relevant information, there is always also non-relevant information. For instance, in PRF, some documents in the feedback set might be non-relevant, or in TRF, some documents, despite the fact that they are relevant, may act like poison pills [38] by hurting the performance of feedback systems, since they also contain off-topic information. Such non-relevant information can distract the feedback model by adding bad expansion terms, leading to *topic drift* [17, 28]. It has been shown that based on this observation, existing feedback systems are able to improve the performance of the retrieval if feedback documents are not only relevant, but also have a dedicated interest in the topic [17]. Given that taking advantage of feedback documents requires a robust and effective method to prevent topic drift caused by accidental, non-relevant terms brought in by broader topic or multiple topics documents in the feedback set.

This paper introduces *significant words language models* (SWLM) to extract a language model of feedback documents that captures the essential terms representing a *mutual notion of relevance*, i.e. a representation of characteristic terms which are supported by all the feedback documents. The general idea of SWLM is inspired by the early work of Luhn [24], in which he argues that to extract **significant words** by avoiding both common observations and rare observations. More precisely, Luhn assumed that frequency data can be used to measure the significance of words to represent a document. Considering Zipf's Law, he simply devised a counting technique for finding significant words. He specified two cut-offs, an upper and lower (see Figure 1), to exclude non-significant words.

There have been efforts to bring this idea into the feedback systems, like mixture models [46] and parsimonious language mod-

Standard-LM	
prize	5.55e-02
nobel	3.36e-02
physics	2.35e-02
science	2.18e-02
...	
time	1.68e-02
...	
palestinian	1.34e-02
year	1.34e-02
...	

General-LM	
new	3.70e-03
cent	2.98e-03
two	2.97e-03
dollars	2.76e-03
people	2.71e-03
...	
time	2.47e-03
...	
year	2.16e-03
...	

SMM [46]	
prize	6.07e-02
nobel	4.37e-02
awards	3.43e-02
chemistry	3.23e-02
physics	2.82e-02
palestinian	2.18e-02
cesium	2.09e-02
arafat	1.94e-02
university	1.92e-02
...	

Specific-LM	
insulin	2.25e-02
palestinian	2.15e-02
dehmelt	1.81e-02
oscillations	1.79e-02
waxman	1.69e-02
marcus	1.69e-02
attack	1.61e-02
...	
arafat	1.29e-02
...	

SWLM	
prize	6.02e-02
nobel	4.53e-02
science	2.68e-02
award	2.43e-02
physics	1.94e-02
winner	1.90e-02
won	1.80e-02
peace	1.80e-02
discovery	1.71e-02
...	

Figure 2: Extracting *significant terms* from relevant feedback documents. (topic 374 of the TREC Robust04 test collection: "Nobel prize winners")

els [18]. They tried to make the feedback model better by eliminating the effect of common terms from the model. However, instead of using fixed frequency cut-offs, they made use of a more advanced way to do this. Hiemstra et al. stated the following in their paper:

> *[...] our approach bears some resemblance with early work on information retrieval by Luhn, who specifies two word frequency cut-offs, an upper and a lower to exclude non-significant words. The words exceeding the upper cut-off are considered to be common and those below the lower cut-off rare, and therefore not contributing significantly to the content of the document. Unlike Luhn, we do not exclude rare words and we do not have simple frequency cut-offs [...]*

In a way, this paper completes the cycle implementing the vision of Luhn. We introduce a meaningful translation of both specificity and generality against significance in the context of the feedback problem and propose an effective way of establishing a representation consisting of significant words, by *parsimonizing* the feedback model toward not only the common observations, but also the rare observations.

Generally speaking, SWLM is the language model estimated from the set of feedback documents which is "specific" enough to distinguish the features of the feedback documents from other documents by removing general terms, and in the same time, "general" enough to capture all the shared features of feedback documents as the notion of relevance, by excluding document specific terms. To do so, in order to estimate SWLM, it is assumed that terms in the feedback documents are drawn from three models: 1. *General model*, representative of common observations, 2. *Specific model*, representative of partial observations, and 3. *Significant words model* which is a latent model representing the notion of relevance. Then, it tries to extract the latent significant words model as the feedback language model.

Figure 2 shows an example of estimating language models from the set of top seven relevant documents retrieved for topic 374, "Nobel prize winners", of the TREC Robust04 test collection. Terms in each list are selected from top 50 terms of the models estimated after stop word removal. Standard-LM is the language model estimated using MLE considering feedback documents as a single document. SMM is the language model estimated using simple mixture model [46], one of the most powerful feedback approaches, which generally tries to remove background terms from the feedback model. General-LM denotes the probability of terms to be common based on their overall occurrence in the collection and Specific-LM determines the probability of terms to be specific in the feedback set, i.e being frequent in one of the feedback documents but not the others. The way General-LM and Specific-LM are estimated will be discussed in detail ahead. And the last model in the figure is SWLM, which is the extracted latent model with regards to General-LM and Specific-LM, using our proposed approach.

As can be seen, considering feedback documents as a mixture of feedback model and collection model, SMM penalizes some general terms like "time" and "year" by decreasing their probabilities. However, since some frequent words in the feedback set are not frequent in the whole collection, their probabilities are boosted, like "Palestinian" and "Arafat", while they are not good indicators for the whole feedback set. The point is although these terms are frequently observed, they only occur in some feedback documents not most of them, which means that they are in fact "specific" terms, not significant terms. By estimating both general model and specific model and taking them into consideration, SWLM tries to control the contribution of each feedback document in the feedback model, based on its merit, and prevent the estimated model to be affected by indistinct or off-topic terms, resulting in a significant model that reflects the notion of relevance.

The main aim of this paper is to develop an approach to estimate a robust model from a set of documents that captures all, and only, the essential shared commonalities of these documents. Having the task of feedback in information retrieval as the application, we break this down into three concrete research questions:

RQ1 *How to estimate significant words language models for a set of feedback documents capturing a mutual notion of relevance?*

RQ2 *How effective are significant words language models in (pseudo) relevance feedback?*

RQ3 *How do significant words language models prevent the feedback model to be affected by non-relevant terms of non-relevant or partially relevant feedback documents?*

The rest of the paper is structured as follows. First, in Section 2 we review related work. Then, we explain our approach for estimating significant words language models in Section 3. Sections 4, 5,and 6 present the experimental setup, the results of the experiments on the tasks of TRF and PRF, and comprehensive analysis on the robustness of the proposed approach. Finally, Section 7 concludes the paper and discusses extensions as future work.

2. RELATED WORK

In this section, we discuss about related studies in the problem of feedback in information retrieval. First, we talk about different feedback approaches in particular methods in the language modeling framework. Then, after discussing initiatives focusing on the task of TRF and PRF, we will discuss the main challenges of this tasks and some already proposed methods which address them.

It has been shown that there is a limitation on providing increasingly better results for retrieval systems only based on the original query [40]. So, it is crucial to reformulate the search request using terms which reflect the user's information need to improve the performance of the retrieval systems. To address this issue, automatic feedback methods for information retrieval were introduced fifty years ago [34] and have been extensively studied during past decades. As the earliest relevance feedback approach in information

retrieval, the Rocchio method [34] is proposed in the vector processing environments for changing the query vector to be similar to the relevant documents vectors and dissimilar to the non-relevant documents vectors. Later, probabilistic methods were proposed to select expansion terms from feedback documents based on a term weighting approach [32, 39]. With the development of language models, several feedback approaches have been proposed in this framework to improve the query language model [18, 22, 27, 37, 46]. The mixture model [46], is one of the well-known feedback methods in the language modeling framework which empirically performs well. The idea is to extract a discriminative language model of feedback documents by decreasing weights of the background terms. As an extension to this model, the regularized mixture model has been proposed by Tao and Zhai [37] which not only involves the query model in the estimated feedback model but also has document-specific mixing coefficients to let different documents have a different amount of background terms.

In the relevance model (RM) [1, 22] given the query, a model is estimated as a multinomial distribution over terms that indicates the likelihood of each term given the query as the evidence, based on the occurrences of term together with the query terms in the feedback documents. In a comparable study conducted by Zhai and Lafferty [46], it has been shown that RM3 as a variant of the relevance model is one of the best performing methods which is strongly robust. Divergence minimization [46] is also one of the feedback approaches in the language modeling framework which tries to estimate a feedback model which is close to the language model of every feedback document but far from the collection language model as an approximation of the non-relevant language model. This method generates a highly skewed feedback model which makes it unable to perform well. Recently, Lv and Zhai [27] proposed the maximum-entropy divergence minimization model that, by adding an entropy term, regularizes the original divergence minimization model leading to significant improvements in the performance of the original method.

Parsimonious language model [18] is one of the models employed for feedback [19, 20, 29]. Generally, it tries to describe the feedback model using fewer number of parameters and similar to the mixture model, the common words in the collection are removed from the model in the estimation process which leads to a more lean and mean language model. Zamani et al. [44] considering the feedback problem as a recommendation problem, made use of matrix factorization in order for predicting expansion terms in a weighted manner. There are also some research that employ the similarity of distributed representation of terms as semantic similarity to improve the performance of feedback [30, 31, 43].

Besides the ad hoc studies, there have been some initiatives with the aim of investigation and study the problem of (pseudo-)relevance feedback in detail. In 2003, Reliable Information Access (RIA) Workshop [14, 42] was organized with the goal of understanding the contributions of both system variability factors and topic variability factors to the overall retrieval variability in feedback. Later on in 2008, the Relevance Feedback track was intended as one the TREC tracks and it has been continued for two more years. The initial goal of the TREC Relevance Feedback track was evaluating and comparing different feedback methods [2]. In the next years, besides the comparison of different methods, they tried to investigate some properties of documents and how they affect the relevance feedback performance. More precisely, the tasks focus on studying the notion of what is a good document for relevance feedback and how a system can recognize a good document [3]. In addition to the Relevance Feedback track, the Robust track in TREC defined one of the goals to improve the consistency of feedback systems by focusing on the poorly performing topics [41].

Applying feedback deteriorates the performance of retrieval in some topics, especially in pseudo relevance feedback in which the performance of the feedback run strongly depends on the quality of the top documents in the initial run [6, 14]. On the other hand, using some documents (even relevant documents) might harm the feedback performance [26, 38]. Hence, there are some challenges in the feedback problem like how to determine whether applying the feedback improves the performance for a specific topic, and how to measure the quality of each feedback document and how to incorporate this information in the feedback process. He and Ounis [16] proposed to examine the interests of feedback documents to the query topic using Entropy, which estimates the distribution of query terms in the feedback documents to see to which degree the feedback documents are interested in the topic. In other work [17] they try to detect good feedback documents in PRF by grouping documents employing some features like the probability of the query terms in the feedback document, the similarity of each feedback documents with other feedback documents, and closeness of expansion terms to the regional query terms. Tao and Zhai [36] proposed a two-stage mixture model in which taking the query as a relevant prior, feedback documents are divided into relevant and background documents and only the documents in the relevant group are employed for updating the query model. Collins-Thompson and Callan [7] tried to model feedback uncertainty to improve the robustness. They proposed to perform sampling over the feedback documents as well as the query to generate different sets of feedback documents and several query variants. Then, combining different feedback models from alternative sets, the robustness of the feedback model can be improved.

Arguably, the key issue in the feedback is robustness in terms of being able to deal with non-relevant terms from non-relevant or partially relevant documents. Our proposed approach addresses the robustness problem head on. This is achieved by using the information from the collection and other feedback documents to control the contribution of documents in the feedback model regarding their merit, and to avoid the selection of non-relevant expansion terms.

3. Significant Words Language Models

In this section, we address our first research question: "How to estimate significant words language models for a set of feedback documents capturing a mutual notion of relevance?" First, we briefly discuss feedback in language modeling, then, we explain how SWLM is estimated in detail.

3.1 Feedback in Language Models

Language modeling is a powerful framework used for information retrieval in which the user information need is represented by query language model, θ_q that is typically estimated based on the original query using MLE: $p(t|\theta_q) = {}^{c(t,q)}/_{|q|}$, where $c(t,q)$ is the frequency of term t in q and $|q|$ is the total number of terms in the query. Then, usually having the smoothed language model of documents, the KL-divergence retrieval model [21] is employed to score documents based on the negative KL-divergence between the estimated language models of the query and each document document:

$$Score(d,q) = D(\theta_q||\theta_d) \tag{1}$$

In order to employ relevance feedback in language modeling framework, a feedback language model, $\theta_{\mathcal{F}}$, is estimated using the set of feedback documents and then this model is employed to

expand the query. A common approach for expanding the query is interpolating $\theta_{\mathcal{F}}$ with the original query model [1, 46]:

$$p(t|\theta'_q) = (1-\alpha)p(t|\theta_q) + \alpha p(t|\theta_{\mathcal{F}}), \quad (2)$$

where α controls the amount of feedback. Thereafter the expanded query model is used in Equation 1 for ranking the documents.

The main goal of different feedback approaches is to estimate an effective feedback model, $\theta_{\mathcal{F}}$, from the set of feedback documents. In the next section, we explain how to estimate significant words language models as a proper model for representing feedback documents and we show that using SWLM as the feedback model in Equation 2 for expanding the query, improves the performance of retrieval system in the feedback runs.

3.2 Estimating SWLM

In order to estimate significant words language models, we assume that there are three models from which each document in the feedback set is generated as a mixture sampling from these models: *significant words* model, *general* model, and *specific* model. The significant words model represents the latent model that is desirable to be employed for query expansion (i.e. feedback model) and is a distribution of terms reflecting the notion of relevance. However, the general model and specific model do not necessarily represent topic-centric models. In a way, they are supposed to represent the distribution of terms that are not considered as relevant information. To extract these two models, patterns of the occurrences of terms in different documents are taken into consideration. In loose terms, the general model represents common observed terms and the specific model represents the partially observed terms, which we assume as two different patterns of distribution of non-significant terms.

Each model is represented using a terms distribution, or a unigram language model, θ_{sw}, θ_g, and θ_s. Based on the generative model, each term in a feedback document is generated by sampling from a mixture of these three models independently. Thus, the probability of appearance of the term t in the document d is as follows:

$$p(t|d) = \lambda_{d,sw}p(t|\theta_{sw}) + \lambda_{d,g}p(t|\theta_g) + \lambda_{d,s}p(t|\theta_s), \quad (3)$$

where $\lambda_{d,x}$ stands for $p(\theta_x|d)$ which is the probability of choosing the model θ_x given the document d.

Based on the patterns of term occurrences in the documents as external knowledge, we estimate θ_g and θ_s and make them fixed in the estimation process as infinitely strong priors. We consider the collection model, θ_C as an estimation for θ_g:

$$p(t|\theta_g) = p(t|\theta_C) = \frac{c(t,C)}{\sum_{t' \in V} c(t',C)}, \quad (4)$$

where $c(t,C)$ is the frequency of term t in the collection. This way, terms that are well explained in the collection model get high probability and are considered as general terms.

Furthermore, we define specificity in the context of feedback problem as being supported by part of the feedback documents but not all. We estimate θ_s to represent the probability of a term being partially observed as follows, and normalize all the probabilities using Softmax normalization, to recover the probability values and establish a well-formed distribution:

$$p(t|\theta_s) \xleftarrow[Normalization]{Softmax} \sum_{d_i \in \mathcal{F}} \left(p(t|\theta_{d_i}) \prod_{\substack{d_j \in \mathcal{F} \\ j \neq i}} (1 - p(t|\theta_{d_j})) \right), \quad (5)$$

where $P(t|\theta_{d_i}) = {}^{c(t,d_i)}/_{\sum_{t' \in d_i} c(t',d_i)}$. Intuitively, Equation 5 calculates the probability of term t to be a specific term. To this end, it considers the probability of a term to be important in one of the document models but not others, marginalizing over all feedback documents. This way, terms that are well explained in only one feedback document but not others get higher probabilities and are considered as insignificant specific terms.

Having the above assumptions, the goal is to fit the log-likelihood model of generating all terms in the feedback documents to discover the term distribution of the significant words model, θ_{sw}. Let $\mathcal{F} = \{d_1, \ldots, d_{\mathcal{F}}\}$ be the set of feedback documents. The log-likelihood function for the entire set of feedback documents is:

$$\log p(\mathcal{F}|\Upsilon) = \sum_{d \in \mathcal{F}} \sum_{t \in V} c(t,d) \log \Big(\sum_{x \in \{sw,g,s\}} \lambda_{d,x} p(t|\theta_x) \Big), \quad (6)$$

where $c(t,d)$ is the frequency of the term t in the document d, and Υ determines the set of all parameters that should be estimated, $\Upsilon = \{\lambda_{d,sw}, \lambda_{d,g}, \lambda_{d,s}\}_{d \in \mathcal{F}} \cup \{\theta_{sw}\}$.

To fit our model, we estimate the parameters using the maximum likelihood (ML) estimator. Therefore, assuming that documents are represented by a multinomial distribution over the terms, we solve the following problem:

$$\Upsilon^* \coloneqq \operatorname*{argmax}_{\Upsilon} p(\mathcal{F}|\Upsilon) \quad (7)$$

Assuming that $X_{d,t} \in \{sw, g, s\}$ is a hidden variable indicating which model has been used to generate the term t in the document d, we can compute the parameters using the Expectation-Maximization (EM) algorithm. The stages of the EM algorithm are as follows:

E-Step

$$p(X_{d,t} = x) = \frac{p(\theta_x|d)p(t|\theta_x)}{\sum_{x' \in \{sw,g,s\}} p(\theta_{x'}|d)p(t|\theta_{x'})} \quad (8)$$

M-Step

$$p(t|\theta_{sw}) = \frac{\sum_{d \in \mathcal{F}} c(t,d) p(X_{d,t} = sw)}{\sum_{t' \in V} \sum_{d \in \mathcal{F}} c(t',d) p(X_{d,t'} = sw)} \quad (9)$$

$$\lambda_{d,x} = p(\theta_x|d) = \frac{\sum_{t \in V} c(t,d) p(X_{d,t} = x)}{\sum_{x' \in \{sw,g,s\}} \sum_{t \in V} c(t,d) p(X_{d,t} = x')} \quad (10)$$

It is noteworthy that estimating general and specific models (θ_g and θ_s) in advance and assuming them as the fixed priors in the EM algorithm, not only helps the EM converge fast (in average less than 100 iterations in our experiments) and the whole procedure to be efficient, but also makes the significant words model θ_s becomes more rigid and accurate and reduces the number of local optimums for λ of different models.

As explained above, besides removing common terms by advocating terms that are relatively rare in the collection, the main contribution of our proposed approach is that it eliminates specific terms by favoring terms occurring in all the feedback documents, not only some of them. There are some previously proposed methods that do this to some degree implicitly. For example by scoring a term based on the multiplication or summation of its probabilities in different feedback documents [22], which will be high if it occurs frequently and evenly in all of them, or by considering the portion of feedback documents that have the term [33], which will be high if all of them have the term. However, in these methods the high frequency of a term in a small portion of feedback documents may compensate a low frequency in others. In our approach the frequency of a term is a privilege, and unless the term frequency is supported by almost all documents the term will be penalized because of its low prevalence.

As a toy example to better understand this, consider we have a set of feedback documents with similar RSV, $\mathcal{F} = d_1, d_2, d_3$, that are

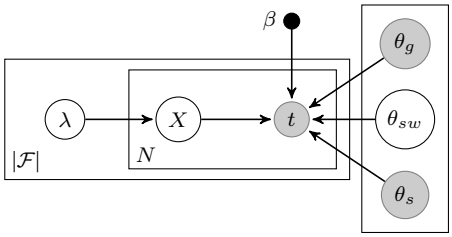


Figure 3: Plate diagram of RSWLM.

all and only relevant documents for the given query, and consider we have two terms t_1, where $\{p(t_1|d_i \in \mathcal{F})\} = \{0.1, 0.1, 0.1\}$ and t_2, where $\{p(t_2|d_i \in \mathcal{F})\} = \{0.02, 0.02, 0.5\}$. Also assume that both t_1 and t_2 have same DF, as well as same TF in the query and the overall collection. In this case, previous methods give a higher score to t_2 compared to t_1, while our approach penalizes t_2 more than t_1, because it has a very high probability in one document and it is not well supported by other feedback documents, hence t_2 probably is a document specific term.

3.3 Regularized SWLM

In the significant words language models, the original query model has not been considered for estimating the feedback model. Thus, in case of having a few relevant documents in the feedback set for a query, the model could be distracted by non-relevant information and converge to a local optimum point. To cope with this problem, and avoid degradation in the performance, a solution is to involve information from the original query [13]. Inspired by the work by Tao and Zhai [37], we modify the estimation process of SWLM and estimate Regularized Significant Words Language Models (RSWLM) by incorporating the extra knowledge from the query model. We define a prior parameter and employ maximum a posteriori to fit the model to feedback documents and solve the following problem:

$$\Upsilon^* \coloneqq \operatorname*{argmax}_{\Upsilon} p(\mathcal{F}|\Upsilon)P(\Upsilon) \tag{11}$$

We define the a conjugate Dirichlet prior on θ_{sw} as follows:

$$p(\theta_{sw}) \propto \prod_{t \in V} p(t|\theta_{sw})^{\beta p(t|\theta_q)}, \tag{12}$$

where $\beta p(t|\theta_q)$ is the parameter of the Dirichlet distribution which in fact performs as the additional pseudo-count for t to push the model θ_{sw} to assign a higher probability to term t as it has a high probability in θ_q. Generally speaking, this adds a bias in the estimation process to bend the feedback model toward the original query model. Here, the value of β controls the amount of the bias. Taking the conjugate prior into account, we conduct the MAP estimation by updating Equation 9 in the EM algorithm as follows:

$$p(t|\theta_{sw}) = \frac{\sum_{d \in \mathcal{F}} c(t,d) p(X_{d,t} = sw) + \beta p(t|\theta_q)}{\sum_{t' \in V} \sum_{d \in \mathcal{F}} c(t',d) p(X_{d,t'} = sw) + \beta} \tag{13}$$

So, by modifying the EM algorithm, we consider our observation from the query model as a pseudo-document which makes the feedback model become more rigid. Similar to the approach in [37], we initialize β with a large value, and then dynamically decrease its value in each EM iteration until the point that we have equal contributions of the original query and the feedback documents. Figure 3 represents the plate notation of regularized significant words language models. As it is shown, for each document the contribution of each of three models, λs, are estimated. It can be seen that general model, θ_g, and specific model, θ_s are considered as external observations, which are involved in the estimation process as infinitely strong priors. It is noteworthy that fixing these parameters also helps to decrease the number of local maximums. As it is illustrated in the diagram, β plays the role of regularizing parameter. The plate diagram of significant words language models would be the same, except there is no regularizing parameter in the model.

In this section, we explained the procedure of estimating SWLM and RSWLM in detail addressing the question "How to estimate significant words language models for a set of feedback documents capturing a mutual notion of relevance?" Establishing a model consisting only the significant words using the fixed cut-offs, as was originally proposed by Luhn [24], runs the risk of leaving good expansion terms out, especially trimming the model toward specific terms may lead to the loss of discriminative relevant terms that can have a high impact on retrieval effectiveness. In our approach, we use the idea of parsimonization which enables us to reduce this risk. On the other hand, estimating specific language model using Equation 5, makes a meaningful translation of specificity against the significance, which empowers our estimation process to retain the significant terms that are globally infrequent, but well supported by the feedback documents.

Table 1: Statistics of the collections used for experimental evaluations

Dataset	Task	Queries	#Docs	#Queries in TRF exp.
Robust04	2004, Robust	301-450 601-700	528,155	217
WT10G	TREC9, 10, Web ad-hoc	451-550	1,692,096	81
GOV2	TREC'04-'06 Terabyte Track	701-850	25,178,548	134

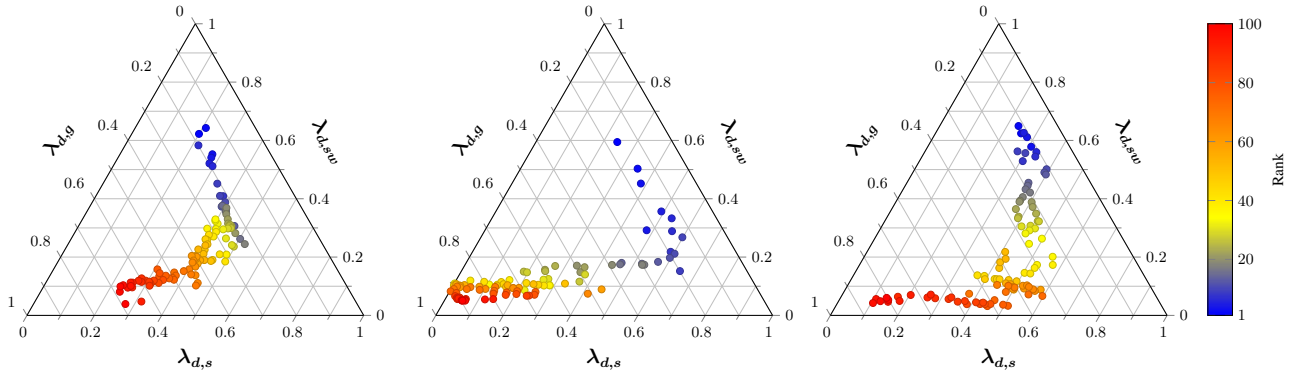
4. EXPERIMENTAL SETUP

In this section, we describe the test collections used in our experiments as well as the settings of our experiments. We use the Robust04, WT10G, and GOV2 test collections, which are different in terms of both size and genre of documents. Information about each collection is summarized in Table 1.

We have employed the Lemur toolkit and Indri[1] search engine to carry out our experiments. We have implemented SWLM and RSWLM in Lemur project framework. In all our experiments, we only use the "title" field of the TREC topics as queries. We have used the Porter stemmer for stemming all queries and document's terms and removed stopwords using the standard InQuery stopword list. We have used the KL-Divergence model [21], with Dirichlet smoothing [45] as the retrieval model in all of the experiments, including initial retrievals as well as feedback runs. We set the Dirichlet smoothing prior to $1,000$. In the feedback runs, for each collection and each method, we have done full grid search and tuned three main parameters: the value of the feedback interpolation coefficient, the number of feedback documents, and the number of expansion terms, by dividing the queries into three folds and conducting 3-fold cross-validation with the same split for folding in all the experiments. Also we have tuned the free parameters of each method during the cross validation.

The Mean Average Precision (MAP) performance measure for top $1,000$ documents is presented as the evaluation metric. Moreover, we report P@10 (precision at 10) for PRF and P@20 (precision at 20) for TRF as the indicators of the precision for the *first* and *two-first* result pages, respectively. To avoid the ranking effect in the evaluation of the TRF task, we have used modified freezing technique in the evaluation of the results of these experiments [5, 13,

[1] http://www.lemurproject.org/



(a) Robust04 dataset **(b)** WT10G dataset **(c)** GOV2 dataset

Figure 4: Contribution of each of the relevance, general, and specific models in the top-100 documents as the feedback set, according to the λs learned in the RSWLM (the average over all the queries).

35]. In addition to the above metrics, we also report robustness index, $RI(Q)$, which is also called reliability of improvement [7]. For a set of queries Q, the RI measure is defined as: $RI(Q) = {}^{N^+-N^-}/_{|Q|}$, where N^+ is the number of queries helped by the feedback method and N^- is the number of queries hurt.

In our experiments, as the baseline methods, we have used the most popular unsupervised state-of-the-arts for the feedback task that are proposed in the language modeling framework. Our baseline methods are: the maximum likelihood estimation—without feedback (MLE) [21], the simple mixture model (SMM) [46], the divergence minimization model (DMM) [46], the relevance models (RM3 and RM4) [1, 22], the regularized mixture model (RMM) [37], and maximum-entropy divergence minimization model (MEDMM) [27].

5. SWLM FOR FEEDBACK

In this section, we investigate our second research question: "How effective are significant words language models in (pseudo) relevance feedback?" We report our experimental results to indicate the effectiveness of significant words language models and regularized significant words language models in the task of pseudo relevance feedback (PRF) and true relevevance feedback (TRF) and compare them to the baseline methods.

5.1 Pseudo Relevance Feedback

Pseudo relevance feedback aims to expand the query to improve the performance of retrieval having no information about the judgements. In PRF, the underlying assumption is that the initial retrieval yields the relevant documents which can be used to refine the query. Thus, assuming the top-ranked documents $\mathcal{F} = \{d_1, \ldots, d_{\mathcal{F}}\}$ from the initial run as relevant, the feedback model $\theta_{\mathcal{F}}$ is estimated and used for the query expansion. Table 2 presents the results of employing significant words language models, regularized significant words language models as the feedback model as well as baseline methods on the task of PRF. As can be seen, RSWLM significantly outperforms *all* the baselines in terms of MAP, in WT10G and GOV2 collections, that are noisy Web collections.[2] Furthermore, it has the highest reliability of improvements in terms of Robustness Index in all the collections. In the PRF task, RSWLM works better than SWLM as it guides the estimator of the feedback model toward the query model and prevents it to be distracted by the noises of non-relevant documents.

[2]Note that we only indicate when (R)SWLM is significantly better than all baseline methods, they are always significantly better than the non-expansion MLE baseline.

Table 2: Performance of different systems on the task of PRF. $^\blacktriangle$ indicates that the improvements over no FB (MLE) and *all* the baseline feedback methods are statistically significant, at the 0.05 level using the paired two-tailed t-test with Bonferroni correction.

Method	Robust04			WT10G			GOV2		
	MAP	*P@10*	*RI*	*MAP*	*P@10*	*RI*	*MAP*	*P@10*	*RI*
MLE	0.2501	0.4253	n/a	0.2058	0.3031	n/a	0.3037	0.5147	n/a
SMM	0.2787	0.4416	0.37	0.2193	0.3264	0.23	0.3214	0.5230	0.41
DMM	0.2701	0.4370	0.31	0.2184	0.3170	0.14	0.3026	0.5211	0.29
RM3	0.2937	0.4683	0.40	0.2406	0.3317	0.26	0.3417	0.5360	0.45
RM4	0.2690	0.4402	0.32	0.2323	0.3273	0.18	0.3316	0.5208	0.37
RMM	0.2681	0.4384	0.28	0.2222	0.3209	0.21	0.3112	0.5193	0.33
MEDMM	**0.2961**	**0.4719**	0.45	0.2413	0.3440	0.25	0.3396	0.5377	0.43
SWLM	0.2918	0.4674	**0.47**	0.2462	0.3377	0.28	0.3423	0.5316	0.50
RSWLM	0.2945	0.4704	**0.47**	**0.2506**$^\blacktriangle$	**0.3427**	**0.31**	**0.3510**$^\blacktriangle$	**0.5419**	**0.53**

Although it has been shown that PRF always improves the average performance of retrieval [14], under some parameter settings, for some topics it decreases the average precision. This is due to the fact that there might be some non-relevant documents in the feedback set containing non-relevant terms resulting to the topic drift in the extracted feedback model [4, 16, 17]. Thus, as one of the main challenging problems in PRF, it is necessary to control the contribution of different feedback documents for inclusion in the feedback model based on their merit [17] for a specific query.

5.2 Relevance Decomposition

Significant words language models empower our proposed feedback method to dynamically determine the quality of each document. Figure 4 addresses the question "How SWLM controls the contribution of feedback documents in the feedback model based on their level of relevancy?" In this figure, as a sample, we take top-100 documents as the feedback set and illustrate the average contribution of each of the significant words, general, and specific models in this documents, according to the λs learned in the regularized significant words language models.

It is an interesting observation that in all the collections the trend of the change in the contribution of three models is similar. In most cases, as the ranking goes down, the contribution of the significant words model decreases, which is in accordance with the relevance probability of documents based on their ranking. However, this decay is slower in the Robust04 dataset compared to WT10G and GOV2 datasets. This is likely because that Robust04 contains newswire articles, which are typically high-quality text data with little noise, in contrast to WT10G and GOV2 which are web collections containing a more heterogeneous set of documents.

Another interesting observation is that in all the collections we

Table 3: Performance of the modified freezing of the results of different systems on the task of TRF. ▲ indicates that the improvements over no FB and *all* the baseline feedback methods are statistically significant, at the 0.05 level using the paired two-tailed t-test with Bonferroni correction.

Method	Robust04			WT10G			GOV2		
	MAP	*P@20*	*RI*	*MAP*	*P@20*	*RI*	*MAP*	*P@20*	*RI*
MLE	0.2725	0.3949	n/a	0.2487	0.3136	n/a	0.3646	0.5318	n/a
SMM	0.3312	0.4829	0.55	0.2582	0.3812	0.31	0.4666	0.5760	0.51
DMM	0.3012	0.4638	0.42	0.2514	0.3609	0.19	0.4219	0.5621	0.42
RM3	0.3411	**0.5001**	0.63	0.3031	0.3849	0.36	0.4717	0.5851	0.55
RM4	0.3241	0.4766	0.37	0.2887	0.3705	0.31	0.4526	0.5781	0.45
RMM	0.3209	0.4719	0.56	0.2873	0.3760	0.34	0.4400	0.5639	0.57
MEDMM	0.3380	0.4891	0.53	0.3140	0.3920	0.34	0.4701	0.5891	0.61
SWLM	**0.3514**▲	0.4920	0.64	0.3155	**0.3976**	0.36	0.4813	**0.6016**▲	0.64
RSWLM	0.3434	0.4911	**0.68**	**0.3277**▲	0.3905	**0.39**	**0.4903**▲	0.5899	**0.69**

see that the top ranked documents are more likely to contain specific non-related terms than general non-related terms. In other words, as the rank of the document increases, the part of the document which is non-relevant becomes more general. One assumption would be that the retrieval models may tend to rank documents with specific non-related terms higher than documents with general non-related terms. However, traditional retrieval models like KL-Divergence do not differ scores of documents based on their non-relevant part. Another hypothesis would be that, the specificity or generality of the non-related parts of documents is a matter of their length. For example, long documents are more probable to have specific non-related terms than short documents. We investigated the length of the retrieved documents based on their ranking in our experiments and, although the retrieval models in general might have some length bias [23], we observed no strong correlation between length and the ranking in our runs.

Based on the observation from Figure 4, we can conclude that the relevant component captured by the significant words dominates the ranking (as would be hoped and expected) and after that the specific component, and lastly the general component (in line with term weighting methods in the ranking models). These observations support that the proposed model is indeed more accurately modeling relevance than standard IR models. More generally, this analysis shows the analytic potential of the proposed model, for example to analyse the ranking of partially relevant or multi-topic documents, based on the generality or specificity of the subtopics involved, which we will defer to future work.

5.3 True Relevance Feedback

True relevance feedback is employed to expand the user query based on either the explicit "relevant"/"non-relevant" judgments given by the user or implicit relevancy information inferred from the user behavior during his interaction with the system, for the top-k results returned by the retrieval system. In our experiments, we simulated this task. We consider the set of relevant documents on the top-10 results (first page of the search engine result page) in the ranked list as the documents judged as relevant by the user to form the feedback set. In our TRF experiments, same as Lv and Zhai [26], we have removed queries that have no relevant documents in their top-10 results from the test collections. Information on the number of queries used for TRF in each collection is given in Table 1.

Table 3 presents the results of different systems on the TRF task. As can be seen, SWLM and RSWLM are best methods in terms of MAP and RI in all the collections and in terms of P@20 in the Web collections. Since we have used the modified freezing technique [5, 35], in the feedback runs, almost the top 10 results are the same as the initial run, so the improvements in the P@20 metric is sort of a reflection for the precision of second 10 results (second page of SERP).

In the TRF task, although the relevancy information of documents are available, since documents can be multi-topic, it is still possible that the feedback mechanism selects the terms from non-relevant parts of the documents. In the Robust04, in which documents are not normally multi-topic, RM3 performs the best in terms of P@20. However, in the Web collections, which is more likely to contain multi-topic documents (partially relevant), SWLM, by controlling the effect of individual documents on the feedback model, significantly outperforms all the baselines.

Unlike the results in Table 2, in which RSWLM performs better than SWLM in terms of all metrics, in TRF, SWLM presents higher performance in terms of P@20. This might be due to the fact that in TRF, there is less noise and consequently less need to lead the feedback model toward the original query model. On the other hand, since RSWLM has no bias to the original query, it has the opportunity to retrieve some documents that are relevant but terms of the original query do not occur in them frequently.

In this section, we presented the results of SWLM and RSWLM in the tasks of PRF and TRF compared to the state-of-the-art methods, in detail addressing the question "How effective are significant words language models in (pseudo) relevance feedback?" We show that the new models are more effective than all previous methods, and also illustrated how they control the contribution of feedback documents in the feedback model based on their merit. Recall that our approach takes a considerable risk by removing specific terms that are the most powerful retrieval cues if relevant, making the feedback task a critical experiment in distinguishing relevant and non-relevant terms. These results provide strong support for effectiveness of significant words language models, and the general intuitions on the importance of building accurate models of relevance underlying them.

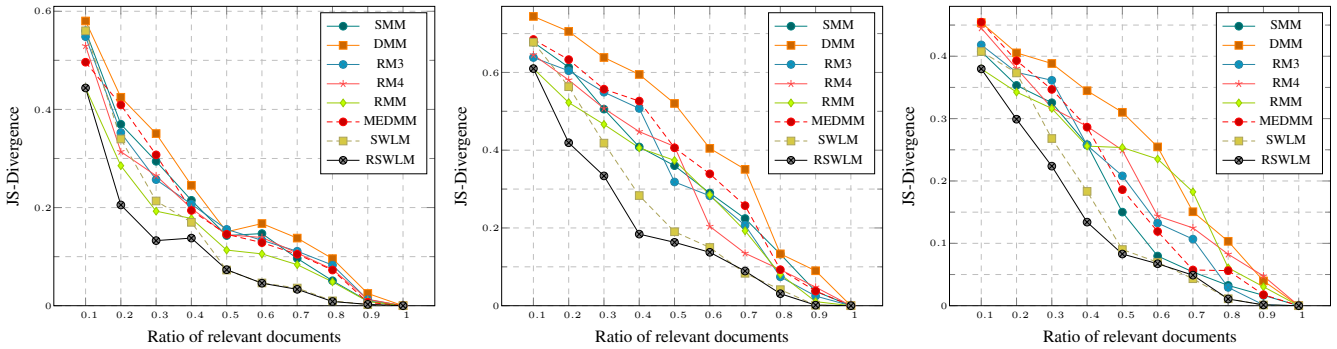
6. ROBUSTNESS

This section investigates our third research question: "How do significant words language models prevent the feedback model to be affected by non-relevant terms of non-relevant or partially relevant feedback documents?" We present analysis resulted from experiments designed to study the robustness of our proposed feedback approach.

6.1 Divergence from Relevance

As an experiment that we have designed to investigate the ability of the each feedback method to deal with noise in the PRF task, using top retrieved results, we measure the divergence of the estimated pseudo relevance feedback models, $\theta_{\mathcal{F}}^{prf}$, from the estimated true relevance models, $\theta_{\mathcal{F}}^{trf}$, that use only those documents from the top retrieved that are explicitly annotated as relevant. In fact, we investigate that how much a feedback method is able to estimate model from a mixture of relevant and non-relevant documents that are similar to the model estimated using only the relevant part. To this end, we calculate the JS-Divergence of $\theta_{\mathcal{F}}^{prf}$ and $\theta_{\mathcal{F}}^{trf}$ for all the approaches and to avoid the effect of the size of the models on the value of divergence, we take top 500 terms of each model as the fixed length representative of the model.

Figure 5 shows the divergence of $\theta_{\mathcal{F}}^{prf}$ and $\theta_{\mathcal{F}}^{trf}$ for different groups of queries with different ratio of relevant documents in top-10 documents, on all collections. As it is expected, in the queries that have a few documents in the top-10 documents are relevant, the divergence is high and two models getting to be the same when all top-10 documents are relevant. In Web collections, convergence of $\theta_{\mathcal{F}}^{prf}$ and $\theta_{\mathcal{F}}^{trf}$ are slower due to the fact that web documents are more noisy and it can be said that usually non-relevant retrieved



(a) Robust04 dataset **(b)** WT10G dataset **(c)** GOV2 dataset

Figure 5: Divergence of true relevance feedback models and pseudo relevance feedback models in different systems, for queries with different ratio of relevant documents in top-10 results.

Table 4: Robustness of different systems against bad relevant documents based on $RI(D_r)$ measure.

Dataset	SMM	DMM	RM3	RM4	RMM	MEDMM	SWLM	RSWLM
Robust04	0.8663	0.7841	0.8716	0.8681	0.8843	0.8914	**0.9319**	**0.9305**
WT10G	0.8504	0.8190	0.8783	0.8961	0.8990	0.9082	**0.9583**	**0.9698**
GOV2	0.8456	0.8062	0.8809	0.8519	0.8910	0.8801	**0.9386**	**0.9209**

documents are farther from relevant retrieved documents, compared to Robust04 dataset. According to the charts in Figure 5, in all collections, SWLM and RSWLM have the least divergence in all the ratios. It means that our proposed models are more robust against being distracted by non-relevant documents. An interesting observation is that in all the collections, the behavior of SWLM and RSWLM are almost the same when at least half of the documents are relevant. In other words, we do not need regularization if at least half of the documents are of interest to the query's topic, either completely or partially.

6.2 Dealing with Poison Pills

Although it has been shown that on average, the overall performance will be improved after feedback [14, 17], for some topics, employing some documents may decrease the average precision of the initial run. As we discussed, in the PRF, it could be due to the fact that the harming feedback documents are not relevant. However, this also could happen in the RF. This is because although the harming feedback document is relevant, there could be only a subset of it containing relevant information. So, adding off-topic terms from this document to the query results in loosing the retrieval performance. These relevant documents that hurt the performance of retrieval after feedback are called "poison pills" [9, 14, 38, 42].

Terra and Warren [38] studied the effect of the poison pills. They used a single relevant document for feedback with several systems to find documents that make the precision drop in all systems. They showed that more than 5% of all relevant documents perform poorly and in one third of all topics there exists at least one bad relevant document which can decrease the performance of the retrieval after relevance feedback.

We have investigated this effect in the multiple feedback documents experiments. In these experiments, for each topic with more than ten relevant documents, we add relevant documents one by one, based on their ranking in the initial run, to the feedback set and keep the track of the change in the performance of the feedback run after adding each relevant document to the feedback set compared to the feedback run without its presence in the feedback set.

To evaluate the robustness of different systems against bad relevant documents, we define a variant of *Robustness Index (RI)* [7] to be applicable in the document level instead of topic level. For a set of relevant documents,D_r, the RI measure is defined as: $RI(D_r) = {}^{N_r^+ - N_r^-}/_{|D_r|}$ where N_r^+ and N_r^- denote number of relevant documents which adding them to the feedback set, based on the above setting, respectively helps or hurts the performance of the feedback run in terms of AP, compared to the case of not including them. $|D_r|$ is total number of tested relevant documents. The higher the value of $RI(D_r)$ is, the more the method is robust against poison pills. Table 4 presents the $RI(D_r)$ of different systems on different datasets. As can be seen, both systems based on significant words language models are strongly robust against the effect of bad relevant documents in all datasets.

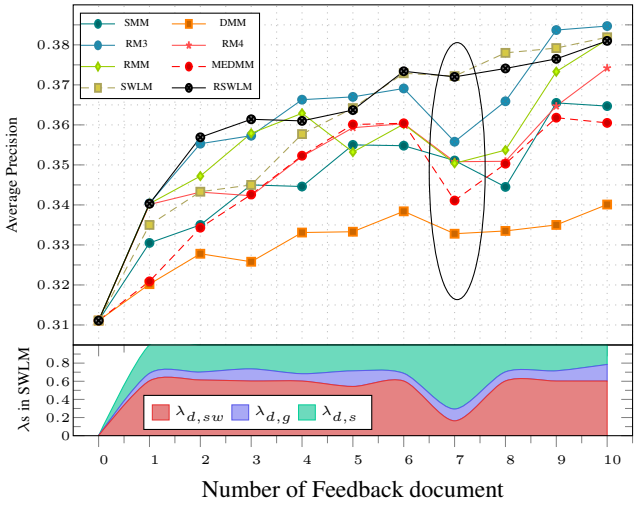


Figure 6: Dealing with poison pills: Effectiveness of different feedback systems facing with a bad relevant document in topic 374 of TREC Robust04.

Furthermore, we have looked into the results of experiments in all the collections and extracted the set of poison pills, i.e. relevant documents that adding them to the feedback set decreases the performance of feedback in *all* the baseline systems. Overal, we found 118 poison pills and we observed that the performance of RSWLM in these situations always has the least drop and in 92% of the cases, it provides the best average precision after adding the poison pill.

As it is discussed by Terra and Warren [38], poison pills are usually relevant documents which have either a broad topic, or several topics. In these situations, employing significant words language models enables the feedback system to control the contribution of these documents and prevents their specific or general terms affect the feedback model. Figure 6 shows how using significant words language model empowers the feedback system to deal with the poison pills. In this figure, the performance of different systems

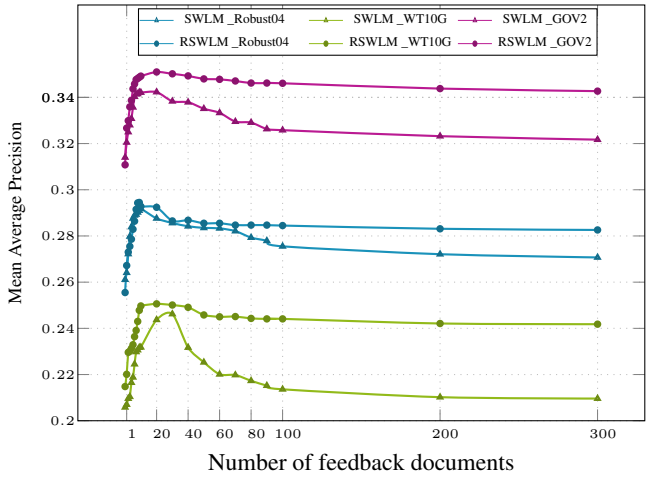


Figure 7: Sensitivity of SWLM and RSWLM to the number of feedback documents

in topic 374 on Robust04 dataset are illustrated. As can be seen, adding the seventh relevant document to the feedback set leads to a substantial decrement in the performance of the feedback in all the systems. The query is "Nobel prize winners" and the seventh document is about one of the Nobel peace prize winners, Yasser Arafat, but at the end, it has a discussion concerning Middle East issues, which contains some highly frequent terms that are non-relevant to the query (see Figure 2). However, RSWLM and SWLM are able to distinguish this document as a poison pill and by reducing its contribution to the feedback model, i.e. learning a low value for $\lambda_{d_7,sw}$, they prevent the severe drop in the feedback performance.

So, our method inoculates the feedback model against poison pills by automatically determining whether adding a specific relevant document to the feedback set hurts the retrieval performance for a specific topic or not and controls its effect in the feedback model.

6.3 Number of Feedback Documents

In order to investigate the sensitivity of our proposed method to the number of documents in the task of PRF, we set all other free parameters to the values that result in optimal average performance and plot the performance of SWLM and RSWLM with regard to the number of documents in the feedback set in Figure 7. As it is noticed, both methods have acceptable robustness. SWLM is more sensitive, especially in Web collections, when low ranked documents are added, it is slightly affected by noises. However, RSWLM is strongly robust and less sensitive to the number of feedback documents.

Furthermore, according to Figure 7, the performance of both systems in all collections is the best when the number of feedback documents are around 10, which is a more or less the same observation in other feedback methods as well [25]. Moreover, this observation is in accordance with the information from the charts in Figure 4, in which the top-10 documents always possess a strong contribution of the significant words model, i.e. high values of $\lambda_{d,sw}$.

In this section, we discussed the robustness of SWLM from different points of view through different experiments, in detail addressing the question "How do significant words language models prevent the feedback model to be affected by non-relevant terms of non-relevant or partially relevant feedback documents?" Results show that SWLM and RSWLM provides robustness against non-relevant or partially relevant documents in PRF, and poison pills in RF. Furthermore, we demonstrate that the performance of SWLM remains stable using different numbers of feedback documents.

7. CONCLUSIONS

This paper concerns the problem of using feedback information to improve the performance of information retrieval. The main aim of this paper was *to develop an approach to estimate a robust model from a set of documents that captures all, and only, the essential shared commonalities of these documents.* Inspired by the discussion on the early work by Luhn [24] about *significant words*, we proposed *significant words language models* (SWLM) of feedback documents by avoiding the distracting effect of common observation as well as rare observation, resulting in models that represent the mutual notion of relevance. The idea is to estimate a feedback model that captures all, and only the essential terms. The model is *specific* enough to distinguish the features of the feedback documents from other documents by removing general terms, and in the same time, *general* enough to capture all the shared features of feedback documents as the notion of relevance, by excluding document specific terms.

Our first research question was: "*How to estimate significant words language models for a set of feedback documents capturing a mutual notion of relevance?*" We proposed an estimation process in which repeatedly the probability of common terms that are already well explained in the collection model are decreased, which removes the effect of general terms. At the same time, the estimation process decreases the weight of terms that are frequent in one of the feedback documents but not others, which removes the effect of specific terms on the model. This way, the final model contains characteristic terms that are also supported by all the feedback documents.

Our second question was: "*How effective are significant words language models in (pseudo) relevance feedback?*" We showed that utilizing significant words language models as the feedback model presents promising performance on both TRF and PRF. Analysing the results, we indicated that the strength of SWLM and RSWLM in feedback is due the fact that they are capable of controlling the contribution of feedback documents in the feedback model based on their level of relevancy which copes with the problem of topic drift in query expansion.

Our third question was: "*How do significant words language models prevent the feedback model to be affected by non-relevant terms of non-relevant or partially relevant feedback documents?*" We assessed the robustness of significant words language models in different experiments. We showed that in PRF, both SWLM and RSWLM provide models that have less noise compared to the models of state-of-the-art methods. Furthermore, we presented results which indicate that compared to the other method, our proposed approaches have the least vulnerability against poison pills (relevant documents that hurt feedback performance) in the task of RF. Moreover, we tested sensitivity of our proposed approaches to the number of feedback documents in PRF and demonstrated that RSWLM effectively deals with several non-relevant documents.

Our general conclusion is that in order to have a feedback model consisting of significant words representing the mutual notion of relevance, it is necessary to not only avoid general terms as done in earlier work, but also specific terms that only represent unique characteristics of each particular observed feedback document. Based on this insight, SWLM presents models of feedback documents in which common and partial observations are left out by careful estimation. Our experiments confirm that this makes our model robust and effective in reflecting the mutual notion of relevance over the set of feedback documents.

The approach of this paper is broadly applicable to mixture models or multiple representations [8], hence can be applied to other kinds of data in different applications like group profiling for content personalization and recommendation [10, 15], multiple document

representations in classification or other poly-representation scenarios, or representing hierarchically structured data [11, 12]. We named our model, significant "words" language model in honor of Luhn, however, it could be employed in non-textual environments, since in general the idea is to extract significant "features" representing the shared essence of a group of objects. Moreover, the SWLM is decomposing the score or 'retrieval status value' of documents into three components: relevant, specific, and general, which provides new handles to investigate retrieval methods in order to better understand the concept of relevance in information retrieval, and provide more accurate unsupervised estimates of the probability of relevance. In turn, having such detailed information available allows us to study the effect of retrieved documents on users behavior in search environments based on their being relevant, specific or general, for example in query reformulation.

Acknowledgments This research is funded in part by Netherlands Organization for Scientific Research through the *Exploratory Political Search* project (ExPoSe, NWO CI # 314.99.108), and by the Digging into Data Challenge through the *Digging Into Linked Parliamentary Data* project (DiLiPaD, NWO Digging into Data # 600.006.014).

REFERENCES

[1] N. Abdul-jaleel, J. Allan, W. B. Croft, O. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. Umass at trec 2004: Novelty and hard. In *TREC-13*, 2004.

[2] C. Buckley and S. Robertson. Relevance feedback track overview: Trec 2008. In *TREC 2008*, 2008.

[3] C. Buckley, M. Lease, and S. M. D. Overview of the trec 2010 relevance feedback track. In *TREC 2010*, 2010.

[4] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50, 2012.

[5] C. Cirillo, Y. Chang, and J. Razon. Evaluation of feedback retrieval using modified freezing, residual collection, and test and control groups. In *Scientific Report No. ISR-16 to the National Science Foundation*. 1969.

[6] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *CIKM '09*, pages 837–846, 2009.

[7] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR '07*, pages 303–310, 2007.

[8] M. Dehghani. Significant words representations of entities. In *SIGIR '16*, pages 1183–1183, 2016.

[9] M. Dehghani, S. Abnar, and J. Kamps. The healing power of poison: Helpful non-relevant documents in feedback. In *CIKM '16*, 2016.

[10] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx. Generalized group profiling for content customization. In *CHIIR '16*, pages 245–248, 2016.

[11] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx. Two-way parsimonious classification models for evolving hierarchies. In *CLEF '16*, 2016.

[12] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx. On horizontal and vertical separation in hierarchical text classification. In *ICTIR'16*, 2016.

[13] D. Harman. Information retrieval. chapter Relevance Feedback and Other Query Modification Techniques, pages 241–263. Prentice-Hall, Inc., 1992.

[14] D. Harman and C. Buckley. Overview of the reliable information access workshop. *Inf. Retr.*, 12(6):615–641, 2009.

[15] S. H. Hashemi, M. Dehghani, and J. Kamps. Parsimonious user and group profiling in venue recommendation. In *TREC 2015*. NIST, 2015.

[16] B. He and I. Ounis. Finding good feedback documents. In *CIKM '09*, pages 2011–2014, 2009.

[17] B. He and I. Ounis. Studying query expansion effectiveness. In *ECIR'09*, pages 611–619, 2009.

[18] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *SIGIR '04*, pages 178–185, 2004.

[19] D. Hiemstra, J. Kamps, R. Kaptein, and R. LI. Parsimonious language models for a terabyte of text. In *TREC 2007*. NIST, 2008.

[20] R. Kaptein, J. Kamps, and D. Hiemstra. The impact of positive, negative and topical relevance feedback. In *TREC 2008*. NIST, 2009.

[21] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01*, pages 111–119, 2001.

[22] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01*, pages 120–127, 2001.

[23] D. E. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Inf. Retr.*, 11(2): 109–138, 2008.

[24] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, 1958.

[25] Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *CIKM '09*, pages 1895–1898, 2009.

[26] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *CIKM '09*, pages 255–264, 2009.

[27] Y. Lv and C. Zhai. Revisiting the divergence minimization feedback model. In *CIKM '14*, pages 1863–1866, 2014.

[28] C. Macdonald and I. Ounis. Expertise drift and query expansion in expert search. In *CIKM '07*, pages 341–350, 2007.

[29] E. Meij, W. Weerkamp, K. Balog, and M. de Rijke. Parsimonious relevance models. In *SIGIR '08*, pages 817–818, 2008.

[30] A. Montazeralghaem, H. Zamani, and A. Shakery. Axiomatic analysis for improving the log-logistic feedback model. In *SIGIR'16*, 2016.

[31] N. Rekabsaz, M. Lupu, and A. Hanbury. Generalizing translation models in the probabilistic relevance framework. In *CIKM '16*, 2016.

[32] S. E. Robertson and K. S. Jones. Relevance weighting of search terms. *JASIS*, 27(3):129–146, 1976.

[33] S. E. Robertson, S. Walker, M. Beaulieu, and P. Willett. Okapi at trec-7: automatic ad hoc, filtering, vlc and interactive track. pages 253–264. NIST, 1999.

[34] J. Rocchio. *Relevance Feedback in Information Retrieval*, pages 313–323. 1971. reprinted from Scientific Report ISR-9, Computation Laboratory, Harvard University, August 1965.

[35] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, 2003.

[36] T. Tao and C. Zhai. A two-stage mixture model for pseudo feedback. In *SIGIR '04*, pages 486–487, 2004.

[37] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *SIGIR '06*, pages 162–169, 2006.

[38] E. Terra and R. Warren. Poison pills: Harmful relevant documents in feedback. In *CIKM '05*, pages 319–320, 2005.

[39] C. Van Rijsbergen, D. Harper, and M. Porter. The selection of good search terms. *IP&M*, 17:77–91, 1981.

[40] C. J. Van Rijsbergen. A new theoretical framework for information retrieval. *SIGIR Forum*, 21(1-2):23–29, 1986.

[41] E. M. Voorhees. Overview of the trec 2003 robust retrieval track. In *TREC 2003*, pages 69–77, 2003.

[42] R. H. Warren and T. Liu. A review of relevance feedback experiments at the 2003 reliable information access (ria) workshop. In *SIGIR '04*, pages 570–571, 2004.

[43] H. Zamani and W. B. Croft. Embedding-based query language models. In *ICTIR'16*, 2016.

[44] H. Zamani, J. Dadashkarimi, A. Shakery, and W. B. Croft. Pseudo-relevance feedback based on matrix factorization. In *CIKM'16*, 2016.

[45] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01*, pages 334–342, 2001.

[46] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01*, pages 403–410, 2001.